文書比較結果の多目的利用

Document examinations by means of comparative tables

株式会社クレステック　情報技術部システムコンサルタント　**楠本　浩二**

PROFILE　1986 年九州工業大学大学院工学研究科修士課程情報工学専攻修了。同年富士ゼロックス株式会社入社。構造化文書の研究開発に従事。2000 年独立。その後、例規管理システム、法令審査支援システム、損害保険・生命保険約款チェックシステム、製造業マニュアルの文書編集・比較・精査システムなどの設計・開発に従事。

k-kusumoto@crestec.co.jp

1　はじめに

文書を比較して差分を表示する目的はいくつかある。典型的な目的としては、新旧文書の差分の提示、差分の正誤確認、文書間にある同一箇所や類似箇所の検索、重複した内容を持つ文書の発見などがある。文書やファイルの比較に利用されている汎用アプリケーションは多いが、通常、オフィス文書の比較には、Microsoft® Word の比較機能を使い、2 つの版の差分を下線や末梢線で表示している。

しかし、画一的な従来の比較による表示だけでは、前述の比較目的それぞれに最適な結果を得ることができず、重要な情報を見落とす可能性がある。そこで目的に合わせて比較方法を変えて実行し、結果をわかりやすく表示するパーソナル比較ツール[1] を開発した。本稿では、法規、例規、約款、業務規程書、特許明細書、メンテナンスマニュアル、製品取扱説明書、企業間の契約書に、この文書比較ツールおよびこれと連携した応用システムを適用した例をいくつか紹介する。

更に、比較結果を利用することによって、文書改訂時の変更漏れ可能性の指摘や文書内または文書間の文章表現のゆれを発見できるなど、改訂後の文書の精査も可能であることを示し、今後の課題について述べる。

2　文書の差分表示

2.1　表示の目的

表 1 に示すように、本稿では文書間の差分を表示する目的を大きく2つに分類した。

2.1.1　差分内容の提示

1 つ目の目的は、文書の変化内容の提示である。このような文書の例は、改訂内容の精査精度が最も求められる法規や例規といった条建て文書である。変更された条項番号および条文の改正前、改正後の内容を正確に提示する。この差分を提示するために、改正規定（改め文）

表 1　差分表示の目的と事例

(1)	差分内容の提示	文書中の「どこ」が「どのように」変わったか、所定の形式に従って文書中の正確な変更箇所と変更内容を提示したい。	・法規、例規、保険約款の改正 ・薬の添付文書の改訂 ・特許明細書の補正
		文書中の「どこ」が「どのように」、「なぜ」変わったかができるだけわかりやすく表示したい。	・製品マニュアル／取扱説明書の改訂 ・社内業務規程書の更新 ・企業間契約書の作成と更新
(2)	差分からの発見	同一箇所、類似箇所の存在を知りたい。 特定できた箇所の詳細な差分を知りたい。	・関係する文書の一貫性確認 ・文章のゆれの確認 ・文章の流用部分の発見
		同一の文書を知り、内容が同等な場合、単一の文書だけを管理したい。 文書情報の保全をしたい。	・変更履歴の有無を確認 ・オフィスにある重複文書の整理 ・文書改竄の検出

や新旧対照表という形式で作成する。条文は改正規定や新旧対照表を基に議会で承認されて施行される。この改正規定や新旧対照表を人手で作成するには、法制執務の知識[2]を必要とし、多大な作成時間を要したため、これらを解決するためのシステム[3]が開発された。この他の差分提示の例としては、保険約款や薬の添付文書における比較表などがある。いずれも所定の形式で差分を提示する。

一方、企業における取引上の契約書や業務マニュアルの差分を担当者間で確認する場合、複数の人間が契約の内容を協議しながら内容を固める必要がある。担当者は、変更箇所の合意をとるために差分を明示し、変更内容の妥当性を確認する。またメンテナンスマニュアルや取扱説明書も改訂があった場合、読み手の立場に立ったわかりやすい形式で差分を提示する。

2.1.2　差分からの発見

２つ目の目的は文書の差分からの発見である。例えば、保険約款には基本となる約款とその特約がある。変更した基本約款の条項と関係する特約条項を比較して差分を表示し、類似した条項や対応する条項を見つけ、その影響を確認する。

複数の人によって執筆されるマニュアルや仕様書では、変更履歴からは詳細がわからない場合、実際の変更内容を知るために文書を比較して差分を確認する。また企業内の情報セキュリティ管理のために同一情報の多重化を回避したいときは、重複文書の発見によって文書の整理を行うことがある。

2.2　本ツールの位置づけ

法規、例規のように、文書の差分の詳細を正確に示す必要があるシステム[3]においては、対話的な編集システムと密に連携し、文書内に記録される編集時の履歴を利用している。これは、改正規定において、法制執務上の厳密な手法が要求され、文書の比較結果を利用しただけでは改正規定の自動作成はできないからである。そのため、編集中の履歴を文書に残す編集システムを利用する必要がある。

一方、比較ツール[1]は、編集システムを介在せずに任意の文書間の内容を逐次比較し、差分を計算する。差分表示からの発見を目的とした時、この比較時の処理方法をいくつか指定可能にすることによって、差分からの発見を容易にする。企業においては、わかりやすい差分の提示と、差分からの発見を要求する業務の場面が多いため、この手法を活用できる。また、本手法においても編集システムを利用したときのように、文書間の相違点と類似点を正確に特定できるようにし、様々な目的に応じた差分提示を可能にすることを目標にしている。

3　比較方法の指定

3.1　比較単位

比較結果の表示には、１つの画面内に変更箇所を下線や末梢線で表示した形式、差分をビジュアルに対応付けた形式、対照表形式などいくつかの形式がある。本ツールでは比較結果の表示として、文書の変更前後で対応する箇所を左右に並べて表示する対照表形式を採用した。この表示形式の採用理由は以下である。

・対応する箇所が表の一行として横並びになるのでわかりやすい
・第三者に提出する公式文書への整形が容易
・表形式の場合、目的に従ったフィルタリング、ソートが可能なので比較結果を見やすい形式に加工しやすい

対照表として左右に並べるとき、例えば、見出しと本文との比較、段落と表との比較は意味がない。そこで対応する箇所の区分けを指定することによって、文書の内容を所定の単位に分解する。この単位は通常、段落、文、表、それ以外の要素、例えば画像である。文書が XML のような構造化された文書形式の場合には要素名や属性情報も手掛かりに分解できる。法規や例規といった文書

表２　比較単位

比較単位	比較対象
段落	改行を終端とする文字列
文	段落内を更に区切り、和文は句点「。」英文はピリオド「.」を終端とする文字列
表	表構造下の行を単位とする文字列
条	〈条〉〈項〉…〈/項〉〈/条〉中の文字列
文字列以外	例えば画像など

は条建てなので、条または項を比較単位として指定すると意味のある要素同士の比較が可能になる。

3.2 比較方法

本ツールは、以下の2通りの比較方法を提供している。目的に従っていずれかを指定する。第1の比較方法では、文書A中のm個の比較対照の要素を順序付リストA$\langle a_1, a_2, a_3, \cdots, a_m \rangle$と表し、文書B中のn個の要素を順序付リストB$\langle b_1, b_2, b_3, \cdots b_n \rangle$と表す。この比較方法は、文書Aと文書Bの相対する要素間の類似度similarity(a_1, b_1)、similarity(a_2, b_2)、similarity$(a_3, b_3)\cdots$を先頭から順に求めていく比較である。したがって比較の回数はmax(m, n)となる。

第2の比較方法では、文書Aを基軸として文書Bの要素全体を比較対象の集合と捉え、類似性が最も高い要素を検索する総当たりの比較をする。例えば、a_1に関する類似度はmax(similarity (a_1, b_i)) $(i=1, \cdots, n)$となる。したがって比較の回数は$m \times n$となる。この場合も類似度が一定の閾値以上の場合、その差分を対照表の同一行に表示する。この比較方法によると、比較回数が多いために実行時間はかかるが、文書Aと文書Bにおいて、文書の論理構成が異なるために要素の出現順序が全く異なっていても差分を正確に表示できる利点がある。

3.3 類似度算出と閾値

本ツールでは、類似度と同時に差分を求める必要があるため、ルーベンシュタインの編集距離[4]を適用した。任意の文字列String1とString2の類似度similarity(String1, String2)の算出式を式1に示す。

$$\frac{\max(\text{length}(\text{String1}),\ \text{length}(\text{String2})) - (\text{String1}\text{と}\text{String2}\text{の編集距離})}{\max(\text{length}(\text{String1}),\ \text{length}(\text{String2}))} \times 100$$

式1　類似度の計算式

この値からString1とString2との関係をデフォルトとして表3のように判定する。

なお、このデフォルトの類似度閾値はユーザーが任意の値に変更できる。

表3　類似度閾値と比較結果の判定

	和文	英文	判定
類似度閾値	30%未満	40%未満	無関係
	30%以上	40%以上	変更
	60%以上	60%以上	類似／ゆれ
	100%	100%	一致

3.4 語句単位の編集距離

文字列間の編集距離を求める際、1文字単位ではなく形態素単位での編集距離とすると、例えば文字列「電気回路」から「電子回路」への差分があったとしても「気」⇒「子」といった一文字の差分ではなく「電気」⇒「電子」として処理できる。更に連続する名詞の列を1つの語句として差分とすると「電気回路」⇒「電子回路」として差分表示できる。このような語句は、形態素の所定の出現パターンを定義しておくことによって識別できる[1]。

3.5 比較結果の表示

表4は、和文特許明細書の一部を比較した結果の一例である。右の数値は、旧文書と新文書の文字列に対する類似度を表示している。この例では85%の類似度なので表3より類似箇所と判定して同一行に表示し、差分箇所を太字かつ下線で示している。類似度が30%未満だった場合は、関連しない別の内容であると判定して同一行に表示しない。

表4　和文の比較結果

旧	新	類似度
【請求項**2**】　表示**装置**と制御**装置**を含む請求項1記載のシステム。	【請求項**3**】　表示**部**と制御**部**を含む請求項1記載のシステム。	85%

上記の和文を英文翻訳にした場合の類似度は表5になる。英文の場合、文字の区切りにスペースが存在するため、これらが一致して和文の見かけの類似度よりも高い数値が出る。したがって英文の場合、比較時に空白文字を無視し、単語単位で比較するように調整する。なお和文と英文とは比較しない。

表5　英文の比較結果

旧	新	類似度
2. The system of Claim 1 comprising a display **device** and a control **device**.	**3**. The system of Claim 1 comprising a display **part** and a control **part**.	81.25%

4　比較結果の利用

4.1　差分からの発見

以上の比較方法に従って文書を比較すると、特に新旧文書だけではなく、相互の関係がわかっていない任意の文書間の差分発見を目的とすることができる。

4.1.1　文書要素の移動

改訂の前後で文が移動した結果を表示する例を表6に示す。以下、本稿では説明のために対照表内の数字は、比較単位である文や段落の識別子を示している。

旧文書の1つ目の文と新文書の3つ目の文の類似度が100% であるにも関わらず、要素の出現位置だけが異なる場合、要素の移動を表示する。この例では、記載内容自体に変化はなかったが、最初の文が最後に移動したことを検知できる。

表6　要素の移動を表示

旧		新	
1	本発明の目的は、攻撃に強いシステムを構成することである。		（3へ移動）
2	構成を示すと本システムは暗号化部を有している。	1	構成を示すと本システムは暗号化部を有している。
3	本システムは複号化部を有している。	2	本システムは複号化部を有している。
	（1と同一）	3	本発明の目的は、攻撃に強いシステムを構成することである。

4.1.2　文書要素の順序変更

特許明細書において、段落の出現順に請求項を比較し、表7の結果を得たとする。この結果からは請求項2と請求項3の内容それぞれに変更があったことを表示している。

表7　要素の出現順序を優先した比較結果

旧		新		類似度
2	【請求項2】 **入力**装置と**出力**装置を含む請求項1記載のシステム。	2	【請求項2】 **表示**装置と**制御**装置を含む請求項1記載のシステム。	90%
3	【請求項3】 **表示**装置と**制御**装置を含む請求項1記載のシステム。	3	【請求項3】 **入力**装置と**出力**装置を含む請求項1記載のシステム。	90%

しかし、要素の出現順に関係なく、一旦、全要素と比較し、類似度が高いものを変更箇所とすると表8となる。この例では 90% よりも大きな類似度 95% が求められたため順序を変更している。

この結果から請求項2と請求項3の内容が不変だったことと、これらの請求項の順序変更があったことを検知できる。

表8　類似度を優先した比較結果

旧		新		類似度
2	【請求項**2**】 入力装置と出力装置を含む請求項1記載のシステム。	3	【請求項**3**】 入力装置と出力装置を含む請求項1記載のシステム。	95%
3	【請求項**3**】 表示装置と制御装置を含む請求項1記載のシステム。	2	【請求項**2**】 表示装置と制御装置を含む請求項1記載のシステム。	95%

4.1.3　文書の同一性

文書Aと文書Bの内容が同一か否かは双方のファイルのタイムスタンプや文書サイズを比較する場合がある。しかし、これらが等しくても内容が同一である保証はないため、文書の内容を以下の視点から比較する必要がある。

- 文書内の内容が出現順を含めて完全に一致しているか
- 文書に書かれていることが出現順を問わず、一致しているか

前述したように文書要素の移動や入れ替えの検知ができるので、要素の出現順を問わないような文書の同一性も発見できる。

4.1.4　改訂に伴う変更漏れ

多大な時間をかけて精査された文書もその後、更新される。元の文書が正しくても一旦変更が加えられると、わずかな変更であってもその文書の正確性が失われる。一旦更新された文書は、再度精査をする必要がある。文書の精査にかかるコストは文書サイズに比例して大きくなるため、変更した文書内の追加、削除、変更の箇所だけをチェックするに留めることが多い。表９に保険約款の例で説明する。

表９　参照先の差分を検知

旧	新
１．保険契約者、被保険者が権利を放棄したとき	１．保険契約者、被保険者が権利を放棄したとき
２．この保険契約の付加特約が重大事由によって解除されたとき	（号が削られた）
３．給付金の請求に関し、給付金の受取人に詐欺行為があったとき	**２**．給付金の請求に関し、給付金の受取人に詐欺行為があったとき
４．その他この保険契約を継続することを期待しえない前２号に掲げる理由と同等の理由があるとき	**３**．その他この保険契約を継続することを期待しえない前２号に掲げる理由と同等の理由があるとき
	↑**前２号**の参照先不整合

この例では、改訂によって第２号が削られ、第３号と第４号が繰り上がってそれぞれ第２号と第３号になっている。変更後も第３号の前に２つの号が存在しているため、この号が削られたことによる影響を見落としている。

本ツールでは、この対照表の結果から変更後の第３号の文中にある「前２号」が示す引用先の内容の一部が変更前と一致しないことを検出して「前２号」に対する変更時の修正漏れの可能性を指摘する。この指摘を見たユーザーは、前２号の参照先である第２号が削られたので、この変更に伴い、「前２号」を「前号」に訂正する必要があることに気づく。

このように、改訂前後の文書を比較することによって、単一の文書だけでは発見が困難な参照関係のずれを発見でき、改正後の文書のチェックにかかる時間を節約できる。

4.1.5　表現のゆれ

本ツールでは、１つの文書内において文書要素同士を比較することによって、その文書内の「表現のゆれ」を検出する。この手法を利用すると、同一か同等であるべきなのに異なる表現になっている箇所やタイプミス、誤字脱字も発見できる。

しかし、全数比較は性能面で課題がある。文書中の比較対象要素の数が n 個の場合、比較の回数は $n \times (n-1)/2$ となるので、文書サイズが m 倍になると m^2 以上の時間が必要となる。そこで要素同士の不要な比較を回避するために、文と文以外の段落とを区別する。文以外の段落としては、見出し、品目、箇条書の項目などがある。文章以外の要素も多いマニュアルや取扱説明書では、文同士の比較だけを実行することによって高速化できる。図１に手順を示す。

STEP1：文書内の「文」と「文以外の段落」を判定
STEP2：「文」と「文」and/or「文以外の段落」と「文以外の段落」の組み合わせすべてに関して類似度および差分を計算
STEP3：所定以上の高い類似度を有する場合、対照表の同一行に差分を表示

図１　「表現のゆれ」検出の手順

ある製品の取扱説明書３冊の合計 970 頁を対象に類似度閾値 60%で比較を実行したところ、重複を除いた 275 例を検出した。文の総数は 7431 個であった。文同士の差分結果から表 10 のように分類した。一方、この文書において、文以外の段落間の比較を実行したところ、ほとんど意味のない結果となった。

表 10　類似度 60% 以上を有する文同士の差分の内訳

表現のゆれ	名詞に差異	読点の有無記号に差異	助詞のゆれ	誤字・脱字
137 例	46 例	45 例	41 例	6 例
49.8%	16.7%	16.4%	14.9%	2.2%

上記分類中、最も多かった「表現のゆれ」の一例を図２に示す。同一の内容なのに、具体的な表現、簡単な表現、一貫性がない表現などがあることを発見できる。

Word には「表記ゆれチェック」の機能があるが、図２にあげた例のどれも検出されない。ここでリストされた表現のゆれに含まれているパターンの一部を以下に示す。

・類似した名詞や動詞の使用

・副詞や修飾語の有無

・語尾の差異

文書が人による著作物である以上、表現のゆれは必然

キーを**押すと**基本画面**が**表示さ**れ**ます。 キーを**押し**基本画面**を**表示さ**せ**ます。
機器 v 始動**後**、始動スイッチを**放し**てください。 機器**が**始動**するまでは**、始動スイッチを**押し続け**てください。
もし、不具合が**継続する**場合は、最寄りの弊社支店または営業所に**お問い合わせ**ください。 v 不具合が**解消されない**場合は、最寄りの弊社支店または営業所に**ご相談**ください。
スイッチを v 時計**方向**に回すと風量が**強くなり**ます。 スイッチを**反**時計**回り**に回すと風量が**弱まり**ます。
右に回すと**周波数が高く**なり、左に回すと**周波数が低く**なります。 右に回すと**高音に**なり、左に回すと**低音に**なります。
レバーを**確実に引き上げて**「ロック位置」にし**てください**。 レバーを v「ロック位置」にし**ます**。

図２ 「表現のゆれ」の一例

的に発生しうる。特に、複数の人によって時間をかけて作成する製品マニュアルを本ツールで精査すると、共同執筆者による書き方の相違、校正担当者の評価基準の相違、時間経過に伴う人の書き方の変化、などが発生していることがわかる。

特許明細書において、特許ライティングマニュアル[5]の第Ｆ条の２にある「文レベルの表現揃え」にあるような文が複数あった場合、１箇所でも異なる文を表現のゆれとして発見できる。特許明細書ではこのような表現のゆれが、権利範囲に影響を及ぼす可能性もある。また機械翻訳、翻訳メモリーによる翻訳では、予め原文の表現のゆれをなくすことが望ましい。

4.2 差分の提示

差分を提示する目的のために、本ツールは以下を特徴とする機能を追加している。

4.2.1 新旧対照表の自動生成

新旧文書の間で文章の構成に変更があった文書を比較し、新旧対照表を自動生成する例を紹介する。旧文書では、１つの項目（号）が、新版では２つの項目（号）に分割されたときに自動生成される対照表を表 11 に示す。対照表を一見すると、大きな変更があった比較結果となる。

そこで表 12 のように新文書の第３号を第２号と結合すると、変更内容が更にわかりやすい対照表の作成ができる。この比較結果の場合、新文書には具体的な文言だけが追加されたことを提示できる。

このように自動生成された新旧対照表の比較単位を後で対話的に変更し、差分を再計算した表示ができる。

表 11 自動生成された新旧対照表

旧	新
② 被保険者が運転資格を持た**ないで被保険自動車を運転している場合、酒に酔った状態で被保険自動車を運転している場合、または麻薬等の影響により正常な運転ができ**ない**おそれがある状態**で被保険自動車を運転している場合	② 被保険者が運転資格を持た v ない v で被保険自動車を運転している場合
	③ 酒に酔った状態もしくは身体に道路交通法施行令で定める程度以上にアルコールを保有する状態で被保険自動車を運転している場合、または麻薬等の影響により正常な運転ができないおそれがある状態で被保険自動車を運転している場合

表 12 対話的に編集して作成された新旧対照表

② 被保険者が運転資格を持たないで被保険自動車を運転している場合、酒に酔った状態 v で被保険自動車を運転している場合、または麻薬等の影響により正常な運転ができないおそれがある状態で被保険自動車を運転している場合	② 被保険者が運転資格を持たないで被保険自動車を運転している場合 ¶③ 酒に酔った状態**もしくは身体に道路交通法施行令で定める程度以上にアルコールを保有する状態**で被保険自動車を運転している場合、または麻薬等の影響により正常な運転ができないおそれがある状態で被保険自動車を運転している場合

4.2.2 改正規定の自動生成

新旧の約款を比較して自動生成された表 11 の対照表をもとにして、監督省庁への提出書類である改正規定を自動生成することもできる。表 11 の結果からは、旧約款中の第２号が変更されたことと、新約款には旧約款には対応しない第３号が存在することが検出される。したがって図３のような文章を自動生成し、それを提示できる。

しかし、どのような変更がされているのかをすばやく発見するには、表 11 や図３よりも人が介在して作成し

た表 12 の差分の方が適していることがわかる。

第２号を次のように改める。
「　被保険者が運転資格を持たないで被保険自動車を運転している場合」
第２号の次に次の１号を加える。
「　酒に酔った状態もしくは身体に道路交通法施行令で定める程度以上にアルコールを保有する状態で被保険自動車を運転している場合、または麻薬等の影響により正常な運転ができないおそれがある状態で被保険自動車を運転している場合」

図３　自動生成された改正規定

5　課題

5.1　比較要素の対応付け

任意の文書の比較には、文書間の対応付けの技術が必要である。これに関する研究として、異なる自治体の例規間の条文対応表作成[6]、保険約款や特約のような基礎書類と消費者向けの文書との対応付け[7] などがある。これらの研究においては、手作業で実施した変更箇所指摘結果と比べると自動的な対応付けすべてを 100% 正しくすることが困難であることと、比較する文章の特徴から最適な類似度を求めるアルゴリズムに関する考察がされている。特許明細書は１頁あたり平均 20 以上の文で構成されているが、メンテナンスマニュアルや製品取扱説明書になると、文字列長が短い文や文章以外の要素が多くなる。文以外の段落要素の比較は、編集距離を基にした類似度の場合、精度が低くなる。例えば、見出しである「構造化された文書の特長」と「構造化文書の特徴」の類似度は 50% となり、表現のゆれとしては検出されない。このような場合、最長共通部分文字列やベクトル空間モデルに基づく類似度の併用も考えられる。類似度を求めるとき、文書の要素ごとにアルゴリズムそれぞれの長所を生かした比較を実行する必要がある。

5.2　表の比較

例えば前述した表３に「中文」の列だけを追加し、表 13 のように変更する。

列の追加は表に対する大きな変更であるが、Word や HTML の表の内部構造が行を基本とする構造で記述されているため、表内の文字列は列方向ではなく、行方向に比較する処理となる。その結果、表３と表 13 を比較した結果の一部は表 14 のようになるが、列の追加の差分表示は表 15 が正しい。

表 13　表３に中文の１列を追加

	和文	中文	英文	判定
類似度閾値	30% 未満	30% 未満	40% 未満	無関係
	30% 以上	35% 以上	40% 以上	変更
	60% 以上	60% 以上	60% 以上	類似／ゆれ
	100%	100%	100%	一致

表 14　表３と表 13 の比較結果の一部

旧		新		
和文	英文	和文	中文	英文
30% 未満	**40**% 未満	30% 未満	**30**% 未満	**40% 未満**
30% 以上	**40**% 以上	30% 以上	**35**% 以上	**40% 以上**
60% 以上	**60**% 以上	60% 以上	**60**% 以上	**60% 以上**
100%	100%	100%	100%	**100%**

表 15　期待される比較結果

旧		新		
和文	英文	和文	**中文**	英文
30% 未満	40% 未満	30% 未満	**30% 未満**	40% 未満
30% 以上	40% 以上	30% 以上	**35% 以上**	40% 以上
60% 以上	60% 以上	60% 以上	**60% 以上**	60% 以上
100%	100%	100%	**100%**	100%

因みに、表 15 のような差分の結果から表の変更内容を説明する改正規定は図４になる。

「

和文
30% 未満
30% 以上
60% 以上
100%

」を「

和文	中文
30% 未満	30% 未満
30% 以上	35% 以上
60% 以上	60% 以上
100%	100%

」に変更。

図４　表の変更を提示した改正規定の例

大きな表を比較し、その差分をわかりやすく表示するときには、以上のように変更された表の一部だけを提示することも考えられる。

6　おわりに

本稿では、任意の文書を比較し、その差分を発見す

ることによって文書の精査が可能であること、更に、第三者へわかりやすい差分の提示が可能であることを述べた。また比較結果を対照表形式で表示すると、文書単独の精査では時間がかかり、見落としがちだった情報の発見や、大容量の文書中の文章の表現ゆれの発見も容易になる利点があることも示した。その一方、文以外の段落や表のように、文書要素によっては、比較の精度に開きが生じている。今後、言語処理技術や文書要素の対応付けの技術を活用し、文書要素の差分を更に正確に表示することによって、様々な形式による差分提示の自動化と、文書精査の効率化を可能にしていきたい。

使用したソフトウェア

(1) 形態素解析エンジン「MeCab」, Ver. 0.99　京都大学情報学研究科—日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト
(2) IPA 品詞体系日本語辞書「IPADIC」, Ver.　2.7.0　奈良先端科学技術大学院大学　松本研究室
(3) 文書比較/日本語精査ツール やまと歌（YAMATO·UTA）, Ver.1.1.0（本稿の例の一部は下記ツールの実行結果を掲載）
http://www.ivysystem.co.jp/yamatoUta/index.html

参考文献

[1] 楠本浩二，山口日緒里，鈴木貴年，千引春菜．特許翻訳の品質を向上するための形態素解析結果を利用した文書比較・日本語精査ツールー歌詠と鶯ーの試作 平成24年度 AAMT/Japio 特許翻訳研究会 第２回特許情報シンポジウム資料集 pp17-24.
[2] 石毛正純『法制執務詳解　新版Ⅱ』ぎょうせい（2012年）.
[3] 齋藤大地，野上正充，鈴木英紀，佐藤正文，高林彰．地方自治体向け例規管理システムの設計と開発 情報処理学会第73回全国大会 2011.
[4] Dan Jurafsky, James H. Martin. “Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”.
[5] Japio 特許情報研究所　特許ライティングマニュアル（初版）「産業日本語」.
[6] 竹中要一，若尾岳志．地方自治体の例規比較に用いる条文対応表の作成支援 自然言語処理　vol. 19, No. 3 pp. 193-212. September 2012.
[7] 丹治広樹，山本和英．保険約款と派生書類の自動対応付け　言語処理学会 第17回年次大会 発表論文集 pp. 868-871 2011年3月.